# Needle Custom Search

## Recall-Oriented Search on the Web Using Semantic Annotations

Rianne Kaptein[1], Gijs Koot[1], Mirjam A.A. Huis in 't Veld[1], and Egon L. van den Broek[2,3]

[1] TNO, Delft, The Netherlands
{rianne.kaptein,gijs.koot,mirjam.huisintveld}@tno.nl
[2] Department of Information and Computing Sciences, Utrecht University, The Netherlands
[3] Human Media Interaction group, University of Twente, The Netherlands
vandenbroek@acm.org

**Abstract** Web search engines are optimized for early precision, which makes it difficult to perform recall-oriented tasks using these search engines. In this article, we present our tool Needle Custom Search[1]. This tool exploits semantic annotations of Web search results and, thereby, increase the efficiency of recall-oriented search tasks. Semantic annotations, such as temporal annotations, named entities, and part-of-speech tags are used to rerank and cluster search result sets.

## 1 Introduction

Generally, Web searches either target homepages or are informational tasks. Both of them can be satisfied with a limited number of search results [4]. Consequently, both the search results and the interface of commercial Web search engines are optimized to excel at these tasks. In practice, only the first 10 or 20 search results are shown on the search results page. Moreover, great care is taken over the presentation of this first page; for example, results from different sources (e.g., news, images, and videos) are included [2]. Their performance can be judged using the quality of the highly ranked search results, which can be measured in terms of early precision and Normalized Discounted Cumulated Gains (NDCG) [12] (cf. [2]).

Analysis of search logs show that between 60% and 85% of searchers view solely the first search result page [4, 9]. In only 4.3% of the sessions do users look at more than three search result pages for a query. So, for most searches, commercial search engines' results are adequate. Nonetheless, there are searches in which users cannot find what they are looking for on one of the first three result pages. In this paper we look at how recall-oriented search tasks on the Web could be better supported.

Forensic investigation is an example of a recall-oriented task. Any detail on any Web page can be important in solving an investigation. Most likely, crucial information is even absent on the 10 most popular pages about the topic of investigation (e.g., a person and an event). So, an exhaustive search is needed to be able to generate an accurate picture. E-discovery is an active field of research because of its use case in the United

[1] The Needle Custom Search tool (NCS) can be accessed at: http://mediaminer.nl/topic_3_context/.

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 750–753, 2014.

States, where e-discovery refers to the requirement that documents and information in electronic form stored in corporate systems are produced as evidence in litigation [7].

The goal of our Needle Custom Search tool (NCS) is to burst the filter bubble, i.e. present users with information which is not biased towards their context or search history, and including novel and diverse search results; not the typical top search results from a large Web search engine [11] (cf. [6]). The tool presented in this article is founded on the principles outlined in [5]. This article will continue with a discussion on the characteristics of recall-oriented search tasks. Subsequently, in Section 3, we present our Needle Custom Search tool. Finally, in Section 4, we present our conclusions.

## 2 Recall-Oriented Search

For the design of our tool, we have taken into account the following characteristics, which can be attributed to recall-oriented search tasks on the Web:

- Query expansion: The typical 2 to 3 word query used to search the Web is not sufficient to ensure the recall of all documents relevant to your search topic [6, 8]. Issuing multiple search queries containing different additions and replacements of search words increases the recall of relevant documents.
- Queries do not always have to be answered instantly. A user is likely to spend a considerable amount of time on the search results; so, often he will be willing to wait a little to get good results back. For example, in e-discovery it is more important to find all relevant documents than to get instant answers [7].
- Often searches will focus on user-generated content. Huge amounts of textual data are generated every day in social networks, blogs, and forums and via tweets. Social media has become part of mainstream e-discovery practice [3]. In contrast, more focused content can be found on (official) homepages and online encyclopedia for tasks such as entity ranking [1].

In this article, we want to investigate how we can utilize existing Web search engines for recall-oriented search. Consequently, on the one hand, we exploit engines' strengths and, on the other hand, we devise workarounds for their weaknesses. The alternative would be to develop a general Web search engine from scratch, optimized for recall-oriented search. Especially for searches in specific domains this should be considered.

## 3 Needle Custom Search (NCS)

Our search tool NCS uses Bing or Google's search APIs to collect Web search results. The URLs are scraped and, subsequently, the pages' textual content is extracted and saved in a database. All text is automatically annotated and, hence, is ready to be reranked and filtered or clustered. An example result page is shown in Figure 1.

Semantic annotations can help in finding a balance between the precision and the recall of the search results. They can significantly contribute to reranking and filtering out search results, which are not relevant in the particular search context [6]. Adding semantic annotations to all search results can help the user to apply a divide and conquer strategy to process the search results. Semantic annotations can be used to cluster

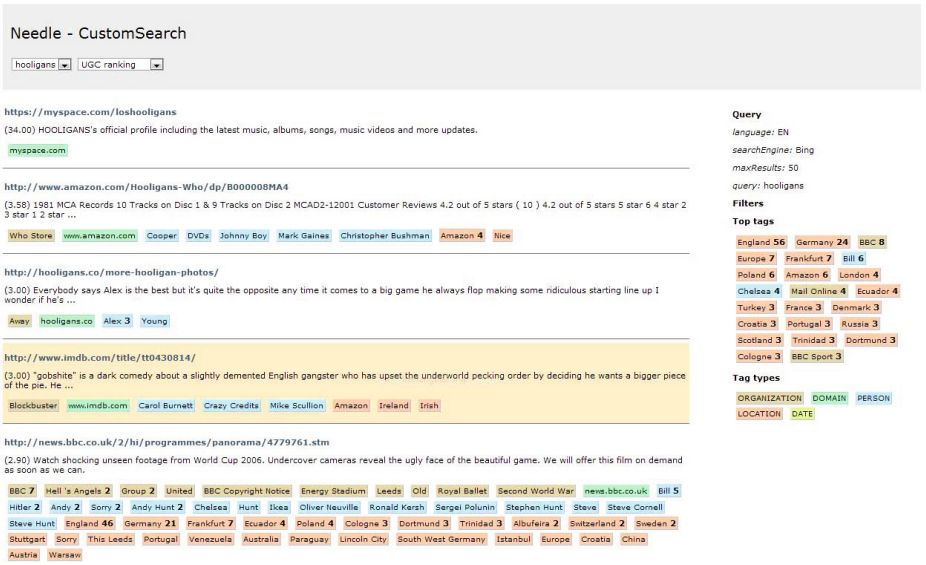


**Fig. 1.** The results page of Needle Custom Search (NCS) for the query 'hooligans'. The search results are ranked by user-generated content probability. For each search result the URL, a snippet, and the assigned tags are shown. Each tag type has its own color, as can be seen in the legend on the bottom right. The most frequently assigned tags are shown. Clicking on any tag will add that tag as a filter on the search results.

or rerank the search results into many dimensions. Using clustering, large number of search results can be processed more quickly by removing irrelevant clusters from your search results, and zooming in on interesting clusters.

NCS is able to make three types of annotations:

- Temporal annotations. Search results can usually be filtered using the date and time a page was last updated. However, many more date indications can be found on the Web pages itself. For example, forum post's date and time stamps.
- Entity type annotations. Named entity taggers extract entities such as persons, locations and organizations from text.
- Part-of-Speech annotations. Part-of-speech tags can tell you something about the type of language that is used. In user-generated content (e.g., a personal blog post), the distribution of part-of-speech tags will be different from those at a commercial shopping page or a Wikipedia page. NCS uses the occurrence of personal and possessive pronouns as an indicator for user-generated content.

NCS uses Heideltime to extract date and time stamps [10] and Apache's OpenNLP library[2] for entity and part-of-speech tagging.

[2] The Apache OpenNLP library. A machine learning based toolkit for natural language text processing; see: `http://opennlp.apache.org/`

The search results, retrieved via Google or Bing's API, can be ranked using:

1. The original ranking as returned by the search API; and
2. The probability that they contain user-generated content; that is, how many personal and possessive pronouns are used.

## 4 Conclusion

We have presented the Needle Custom Search tool (NCS; available online[1]), which supports semantic annotations to enable recall-oriented search on the Web. The aim of this tool is to help users process large numbers of search results more efficiently and prevent them from becoming captured in a filter bubble [6, 11]. For future work we would like to extend the query options to support and encourage longer and faceted queries to improve recall and to visualize the search results in a network instead of a list.

## References

[1] Balog, K., Serdyukov, P., de Vries, A.P.: Overview of the TREC 2011 entity track. In: Proceedings of the Twentieth Text Retrieval Conference, TREC 2011, November 15-18. National Institute of Standards and Technology (NIST), Gaithersburg (2011)
[2] Chuklin, A., Serdyukov, P., de Rijke, M.: Using intent information to model user behavior in diversified search. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 1–13. Springer, Heidelberg (2013)
[3] Gensler, S.: Special rules for social media discovery? Arkansas Law Review 65(7) (2012)
[4] Jansen, B.J., Spink, A.: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management 42(1), 248–263 (2006)
[5] Kaptein, R., Van den Broek, E.L., Koot, G., Huis in 't Veld, M.A.A.: Recall oriented search on the web using semantic annotations. In: Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2013) (2013)
[6] Melucci, M.: Contextual search: A computational framework. Foundations and Trends in Information Retrieval 6(4-5), 257–405 (2012)
[7] Oard, D.W., Webber, W.: Information retrieval for e-discovery. Foundations and Trends in Information Retrieval 7(2-3), 99–237 (2013)
[8] Saracevic, T.: Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. Journal of the American Society for Information Science and Technology 58(13), 1915–1933 (2007)
[9] Silverstein, C., Henzinger, M., Marais, H., Moricz, M.: Analysis of a very large web search engine query log. ACM SIGIR Forum 33(1), 6–12 (1999)
[10] Strötgen, J., Gertz, M.: Multilingual and cross-domain temporal tagging. Language Resources and Evaluation 47(2), 269–298 (2013)
[11] van der Sluis, F., van den Broek, E.L., Glassey, R.J., van Dijk, E.M.A.G., de Jong, F.M.G.: When complexity becomes interesting. Journal of the American Society for Information Science and Technology (in press, 2014)
[12] Wang, Y., Wang, L., Li, Y., He, D., Liu, T.-Y.: A theoretical analysis of NDCG type ranking measures. In: Proceedings of the 26th Conference on Learning Theory (COLT), Princeton, NJ, USA, June 12-14. JMLR: Workshop and Conference Proceedings, vol. 30, pp. 25–54. Microtome Publishing, Brookline (2013)